ЛЕКЦІЯ 3. АНАЛІЗ ЯКОСТІ ОДНОРІДНИХ ПІДМНОЖИН ДАНИХ

НАВЧАЛЬНІ ПИТАННЯ

Відокремлюваність однорідних підмножин даних

Частково відокремлювані підмножини

Повністю відокремлювані підмножини. Методи з обмеженою інформацією.

1. Відокремленість однорідних підмножин даних

Застосування методів порівняльного багатомірного аналізу призводить до розбиття сукупності даних на відносно однорідні підмножини. В залежності від ступеня розбиття на підмножини використовуються різні способи їх оцінювання.

Розбиття на повністю відокремлені підмножини характеризується тим, що об'єкти всередині підмножини слабко розрізняються між собою (тобто, сильно корельовані), а об'єкти різних підмножин мають значні відмінності.

Розбиття на частково відокремлені підмножини характеризується тим, що об'єкти різних підмножин не мають значних відмінностей між собою.

Ступінь відмінності можна визначити за допомогою діаметрів підмножин і відстаней між підмножинами.

Діаметр підмножини P_i визначається за формулою

$$d_{ii} = \max_{p,q \in P_i} d(p,q), \qquad (3.1)$$

де $d(p,q)$ – відстань між елементами підмножини P_i.

Відстань між підмножинами P_i та P_j визначається за формулою

$$d_{ij} = \min_{p \in P_i, q \in P_j} d(p,q). \qquad (3.2)$$

В результаті одержуємо матрицю характеристик розбиття сукупності на однорідні підмножини $D = (d_{ij})$.

Далі визначається граничне значення $d^* = \max d_{ii}$. Значення матриці, що перевищують d^* , до уваги не приймаються, тому що характеризують сильно відокремлені підмножини. Тобто, відмінності між різними елементами цих підмножин більші, ніж відмінності всередині підмножин. Для зручності подальшої обробки даних матриця D перетворюється у матрицю $D¢$ відповідно до зазначеної вище умови

$$d'_{ij} = \begin{cases} d_{ij}, & d_{ij} \le d^* \\ 0, & d_{ij} > d^* \end{cases}. \qquad (3.3)$$

Приклад.

Нехай маємо 10 об'єктів, розташування яких схематично зображено на рис 3.1. В результаті аналізу даних одержано розбиття на об'єктів на три однорідні сукупності: перша – з елементів {1, 2, 3}, друга – з елементів {4, 5, 6, 7}, третя – з елементів {8, 9, 10} . Нехай матриця відстаней між об'єктами має такий вигляд:

$C=$

0	1,2	1,1	3,4	6,1	6,3	5,9	7,4	8,1	12,2
1,2	0	0,8	3,2	4,8	5,7	5,3	6,9	7,7	11,3
1,1	0,8	0	3,6	5,3	5,9	5,6	6,5	7,4	10,8
3,4	3,2	3,6	0	1,2	1,8	1,9	6,4	8,2	9,9
6,1	4,8	5,3	1,2	0	1,4	2,3	6,8	8,7	10,2
6,3	5,7	5,9	1,8	1,4	0	1,5	5,3	6,1	8,2
5,9	5,3	5,6	1,8	2,3	1,5	0	4,2	4,9	7,7
7,4	6,9	6,5	6,4	6,8	5,3	4,2	0	2,2	5,7
8,1	7,7	7,4	8,2	8,7	6,1	4,9	2,2	0	4,8
12,2	11,3	10,8	9,9	10,2	8,2	7,7	5,7	4,8	0

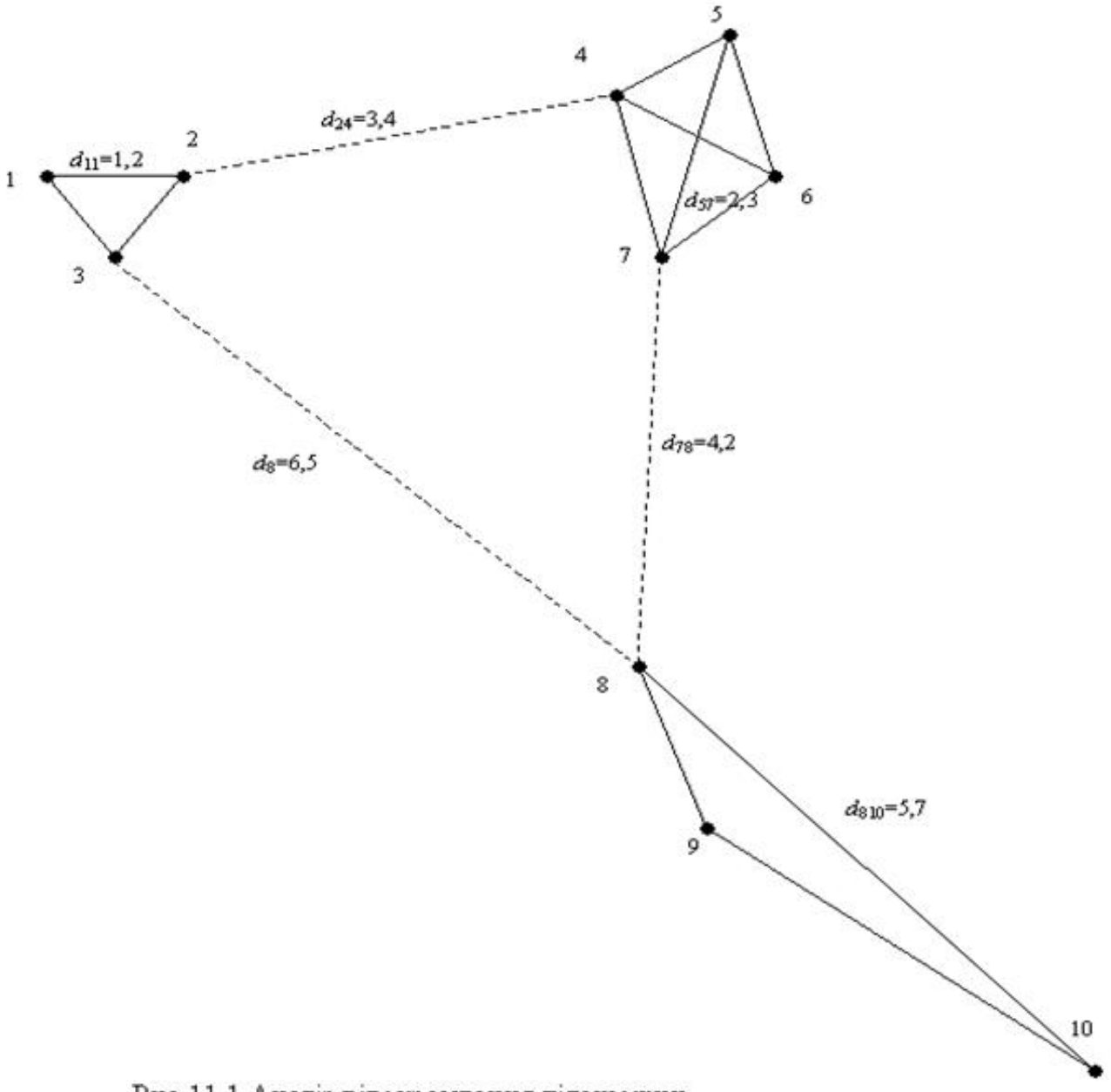

Рис 11.1 Аналіз відокремлення підмножин

За даними цієї матриці побудуємо матриці *D* та *D*¢, які відображають діаметри підмножин та відстань між ними. Для зручності відповідні значення, за якими будувалась матриця *D* відображені у вихідній матриці *C*кольорами. Отже,

$$D=\begin{bmatrix}1,2 & 3,2 & 6,5\\ 3,2 & 2,3 & 4,2\\ 6,5 & 4,2 & 5,7\end{bmatrix}.$$

Скориставшись формулою (3.3) , одержимо матрицю $D¢$ при $d^* = 5,7$:

$$D'=\begin{bmatrix}1,2 & 3,2 & 0\\ 3,2 & 2,3 & 4,2\\ 0 & 4,2 & 5,7\end{bmatrix}$$

Тобто, можна зробити висновок, що перша та третя множина мають хороші характеристики відокремлення. Цей висновок узгоджується з візуальним спостереженням графічного зображення розташування об'єктів. Однак таке зображення може неадекватно представляти істинне розташування об'єктів, оскільки вони можуть характеризуватись багатьма характеристиками, а зображення здійснюється у площині. Висновок про відокремленість інших підмножин та про характеристику подібності об'єктів всередині підмножин зробити поки що не можна.

При розбитті об'єктів на непорожні множини, що неперетинаються, не завжди одержується результат з гарним з точки розу відокремлення підмножин розбиттям. Іноді можна одержати підмножини, подібні між собою та диференційовані всередині. Ці дві особливості підмножин можна описати за допомогою внутрішніх та зовнішніх коефіцієнтів подібності, які обчислюються на основі матриці C відстаней між об'єктами.

Ступінь внутрішньої подібності підмножини P_i визначається за формулою

$$a'_{ii}=\begin{cases}\dfrac{1}{n_i}\sum\limits_{p,q\in P_i}(d^*-d\ (p,q)), & n_i>0\\ d^*, & n_i=0\end{cases}, \qquad (3.4)$$

де n_i –кількість пар елементів підмножини P_i.

Для вимірювання ступеня подібності різних підмножин використовується величина

$$a'_{ij}=\begin{cases}\dfrac{1}{n_{ij}}\sum\limits_{p\in P_i,q\in P_j}(d^*-d^*(p,q))\end{cases}, \qquad (3.5)$$

де

$$d^*(p,q)=\begin{cases}d\ (p,q), & d(p,q)\le d^*\\ d^*, & d(p,q)>d^*\end{cases};$$

n_{ij} – кількість пар елементів підмножин P_i та P_j , для яких виконується умова $d(p,q)$ £ d^*.Високі значення показників свідчать про подібність між об'єктами як всередині підмножини (для a'_{ii}) так і між підмножинами (для a'_{ij}).

В практичному застосуванні такі коефіцієнти незручні, тому їх перетворюють таким чином: $a_{ij}=\dfrac{a'_{ij}}{d^*}$. Очевидно, що 0 £ a_{ij} £ 1. Чим ближче значення коефіцієнта до одиниці, тим вищий ступінь подібності між об'єктами

підмножин.

Приклад

Для розглянутого у попередньому прикладі розбиття об'єктів на підмножини та відповідної матриці відстані між об'єктами *C* одержимо наступну матрицю подібності *A*:

$$A = \begin{bmatrix} 0{,}819 & 0{,}218 & 0 \\ 0{,}218 & 0{,}705 & 0{,}158 \\ 0 & 0{,}158 & 0{,}386 \end{bmatrix}.$$

Як видно, перша та третя підмножина мають низький ступінь подібності. Для елементів першої та другої підмножин спостерігається досить висока подібність всередині підмножин і досить низька – між підмножинами. Найнижча подібність має місце для елементів третьої підмножини. Це говорить про те, що можливо її слід розділити ще на дві підмножини. Це цілком узгоджується з візуальним аналізом зображення об'єктів на рис 3.1.

За значеннями матриці *A* можна побудувати показник, що вимірює ступінь когерентності (зчеплення) елементів різних підмножин. Коефіцієнт простої когерентності має вигляд

$$k_{ij} = \frac{a_{ij}}{a_{ii} + a_{jj}}. \qquad (3.6)$$

Чим більша величина k_{ij}, тим менше відрізняються між собою підмножини. Доповнення цієї величини до одиниці дасть іншу величину – коефіцієнт відокремлення:

$$s_{ij} = 1 - k_{ij}. \qquad (3.7)$$

Являє також інтерес відповідь на питання, в якому ступені кожна з підмножин не схожа на інші підмножини тієї ж сукупності. Для цього служить коефіцієнт індивідуальної когерентності

$$k_{ii} = \sum_{i \neq j} k_{ij} \qquad (3.8)$$

Чим більші значення величини k_{ii} , тим більша схожість даної підмножини з іншими. До цілком відокремлених під множить відносять ті, для яких $k_{ii} = 0$. Інакше підмножини називають частково відокремленими.

Розраховані за формулами (3.6) та (3.8) коефіцієнти утворюють матрицю когерентності *K*. Вона надає можливість оцінити якість розбиття вихідної сукупності даних на підмножини. Це робиться за допомогою узагальненого коефіцієнта когерентності

$$k = \frac{1}{s} \sum_{j>i} \sum_{i} k_{ij}, \qquad (3.9)$$

де *s* – кількість пар підмножин, для яких $k_{ij} > 0$.

Сума береться для елементів матриці *K*, розташованих вище головної діагоналі. По суті, узагальнений коефіцієнт когерентності являє собою середнє значення коефіцієнтів простої когерентності. Він характеризує загальний рівень зближення (або віддалення) всіх підмножин даної сукупності. Чим більше значення коефіцієнта, тим вищий рівень зближення підмножин.

Приклад

Розрахуємо коефіцієнти когерентності для наведених у попередньому прикладі даних.

$$K = \begin{bmatrix} 0{,}143 & 0{,}143 & 0 \\ 0{,}143 & 0{,}288 & 0{,}145 \\ 0 & 0{,}145 & 0{,}145 \end{bmatrix}$$

Отже, цілком відокремлених підмножин в даному випадку немає, але найбільша подібність до інших підмножин сукупності спостерігається для другої підмножини.

Розрахуємо узагальнений коефіцієнт когерентності: $k = 0{,}144$. Таке значення свідчить про досить непогану характеристику розбиття вихідної сукупності даних на підмножини.

1. Частково відокремлені підмножини

Метод транспозиції

Оцінювання параметрів автономних функцій псевдо регресії на основі повної сукупності даних можливе, коли виділені підмножини даних не є повністю відокремленими. В такому випадку оцінювання для окремої підмножини проводиться за даними всіх підмножин, однак спостереження з різних підмножин зважуються:

$$\psi_j = \sum_{i=1}^{m} (y_i - \hat{y}_{ij})^2 \, a_{ij} \to \min, \qquad (3.10)$$

де $\hat{y}_{ij}$ – автономна функція псевдорегресії j-тої підмножини;

a_{ij} – коефіцієнти подібності j-тої підмножини з іншими;

m – кількість спостережень.

Корекція спостережень, віддалених від даних підмножини, для якої проводиться оцінка моделі. Здійснюється за формулами

$$y_i^* = \sqrt{a_{ij}}\, y_i$$

$$x_{is}^* = \sqrt{a_{ij}}\, x_{is}, \qquad (3.11)$$

де Y – пояснювана ознака;

X_S –пояснюючі ознаки , s=1, 2, ..., n .

Величини x_{is} входять в $\hat{y}_{ij}$ – функцію псевдо регресії для j-тої підмножини.

При використанні для корекції даних коефіцієнти простої когерентності використовують формули

$$y_i^* = k_{ij} y_i$$

$$x_{is}^* = k_{ij} x_{is}, \qquad (3.12)$$

$$\psi_j = \sum_{i=1}^{m} (y_i - \hat{y}_{ij})^2 \, k_{ij}^2 \to \min. \qquad (3.13)$$

Метод мультиплікації

Особливістю даного методу є те, що в кожній з виділених підмножин в

декілька раз збільшується кількість спостережень. Збільшення відбувається пропорційно значенням коефіцієнта когерентності:

$$\psi_j = \sum_{g=1}^{k} (10 \cdot k_{ig}^2)^h \sum_{s=1}^{m_j} (y_{sg} - \hat{y}_{sg_j})^2 \to \min \quad , \qquad (3.14)$$

де k – кількість однорідних підмножин даних;

m_j – кількість об'єктів підмножини з номером j;

h – деяка константа, що виражається натуральним числом. Вона визначає значення мультиплікатора 10^h , який показує, у скільки раз збільшується кількість об'єктів у підмножині.

Недолік методу пов'язаний із суб'єктивізмом у виборі значення величини h.

2. Повністю відокремлені підмножини. Методи з обмеженою інформацією.

В тому випадку, коли не виконується одна з головних вимог при оцінюванні параметрів економетричної моделі (кількість значень повинна бути більшою кількості ознак), застосовуються методи з обмеженою інформацією. Ці методи застосовуються також і у випадку розбиття вихідної сукупності даних на повністю відокремлені підмножини.

Перша група методів, що використовуються в такій ситуації, базується на наступному:

1) Сукупність пояснюючих ознак поділяється на підгрупи, щоб кількість спостережень в підмножині була більшою за кількість ознак у підгрупі.
2) Для кожної підгрупи оцінюється функція регресії.
3) Одержані функції агрегується в одну, яка буде містити всі вихідні ознаки.

Оцінювання функції регресії здійснюється окремо для кожної однорідної підмножини даних (тобто, кроки 1)-3) повторюються для кожної підмножини).

Метод блоків, що неперетинаються.

Метод оснований на виконанні наступних операцій:

А) Ознаки розташовуються в наступному порядку: на перше місце ставиться ознака, що має найбільший за модулем коефіцієнт кореляції з пояснюваною ознакою, на друге – ознака з другим за абсолютною величиною коефіцієнтом кореляції і так далі. Далі вихідні ознаки розбиваються на підгрупи. Впорядкування потрібне для того, щоб в кожній з виділених підгруп ознак був приблизно однаковий об'єм інформації: $\sum_{j=1}^{n_i} r_j^2 \approx const$, n_i – кількість ознак у виділеній підгрупі.

Б) Оцінюється незалежно одна від одної функція регресії для кожної підгрупи ознак:

$$\hat{Y}_i = f(X_j) = g_1^1 X_1^1 + g_2^1 X_2^1 + \ldots + g_{n_i}^1 X_{n_i}^1 \quad . \qquad (3.14)$$

В) Обчислюються теоретичні значення $\hat{Y}_j$.

Г) Оцінюються параметри рівняння

$$\hat{Y} = b_1\hat{Y}_1 + b_2\hat{Y}_2 + \ldots + b_k\hat{Y}_k, \qquad (3.15)$$

де k – кількість підмножин.

Метод блоків, що перетинаються.

А) вихідні ознаки впорядковуються аналогічно попередньому методу.

Б) В першу підгрупу відбирається перших n_i ознак.

Кожна наступна підгрупа ознак одержуються шляхом відкидання першої ознаки і приєднання до неї наступної.

В) Для кожної групи ознак оцінюється функція регресії.

Г) Одержані функції регресії агрегується в одну. Оскільки в сусідні $\hat{Y}_i$ входять однакові ознаки X_j , то проводиться їх групування і зведення коефіцієнтів для одержання кінцевого виду рівняння.

Висновки

При оцінюванні параметрів функції регресії в умовах обмеженої інформації застосування різних методів дає взагалі кажучи різні результати. Хоча оцінювані функції містять одні і ті ж пояснюючі ознаки і досить добре узгоджуються з емпіричними даними, значення коефіцієнтів регресії досить часто значно відрізняються. Це можна пояснити тим, що одну і ту ж економічну мету можна досягти різними методами, які еквівалентні з точки зору моделювання, але різні з точки зору структури витрат. Саме в таких випадках доцільно застосовувати методи оцінювання параметрів економетричної моделі в умовах обмеженої інформації.

3. Питання для самоперевірки.

1. В чому сутність проблеми відокремлення підмножин? Які є види відокремлення?
2. Які є характеристики якості розбиття даних на однорідні підмножини? В чому між ними відмінності?
3. Як здійснюється побудова функції регресії для частково відокремлених підмножин?
4. Як здійснюється побудова функції регресії для повністю відокремлених підмножин?